STA221

Neil Montgomery

Last edited: 2017-01-09 13:41

admin

contact, notes

date format	YYYY-MM-DD – *All Hail ISO8601!!!*
instructor	Neil Montgomery
email	`neilm@mie.utoronto.ca`
office	BA8137
office hours	W11-1
website	portal (announcements, grades, suggested exercises, etc.)
github	`https://github.com/sta221-winter-2017` (lecture material, code, etc.)

evaluation, book, tutorials

what	when	how much
midterm 1	2017-02-13	25%
midterm 2	2017-03-27	25%
exam	TBA	50%

I will suggest exercises from the "required" book:

de Veaux, et al., 2014. *Stats: Data and Models*, Second Canadian Edition, 2 edition. ed. Pearson Canada.

Some suggested exercises will be worked out in tutorial each week, starting TBA.

Not a terrible book. I agree with most of it. Too many words. But covers what we need.

Any thick and comprehensive "Stats 101" book could also be a good resource.

software

Data analysis requires a computer. Also, some concepts can be illustrated using simulation, which also requires a computer. We will be using `R`. It's pretty good at data analysis.

language	interpreter	integrated development environment
`R`	`R`	`RStudio`

Some detailed instructions and suggestions for installation and configuration appear on the course website. I will try to impart some data analysis workflow wisdom throughout the course. Some already appears in the detailed instructions.

I have signed up STA221 to have access to some optional R courses on the (well-regarded) datacamp.com training company.

A higher level but thorough resource is:

Grolemund, G., Wickham, H., *R for Data Science* **available free at `http://r4ds.had.co.nz/`**

pre-preliminaries—what is a dataset?

most datasets are rectangles

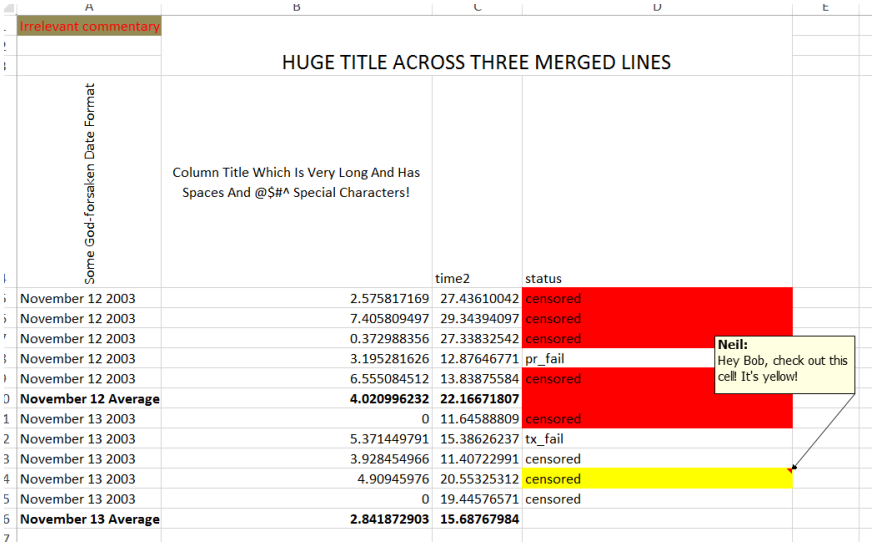
Columns are the *variables*.

The top row has the names of the variables; possibly chosen wisely.

Rows are the *observations* of measurements taken on *units*.

There are no averages, no comments (unless in a "comment" variable), no colors, no formatting, no plots!

not a dataset



	A	B	C	D	E
1	Irrelevant commentary				
2		HUGE TITLE ACROSS THREE MERGED LINES			
3					
4	Some God-forsaken Date Format	Column Title Which Is Very Long And Has Spaces And @$#^ Special Characters!	time2	status	
5	November 12 2003	2.575817169	27.43610042	censored	
6	November 12 2003	7.405809497	29.34394097	censored	
7	November 12 2003	0.372988356	27.33832542	censored	
8	November 12 2003	3.195281626	12.87646771	pr_fail	
9	November 12 2003	6.555084512	13.83875584	censored	
10	**November 12 Average**	**4.020996232**	**22.16671807**		
11	November 13 2003	0	11.64588809	censored	
12	November 13 2003	5.371449791	15.38626237	tx_fail	
13	November 13 2003	3.928454966	11.40722991	censored	
14	November 13 2003	4.90945976	20.55325312	censored	
15	November 13 2003	0	19.44576571	censored	
16	**November 13 Average**	**2.841872903**	**15.68767984**		
17					

Neil:
Hey Bob, check out this cell! It's yellow!

not a dataset

ASSETNUM	MOVEDATE_1	FROM_LOCATION1	TO_LOCATION1	MOVEDATE_2	FROM_LOCATION2	TO_LOCATION2	MOVEDATE_3	FRC
0201011	2005-12-16	NO_LOCATION	RSREPAIR					
0209679	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN4VNCR	2014-02-14	DN4
0209680	2005-05-17	NO_LOCATION	RSREPAIR	2005-08-03	RSREPAIR	WY172UCR	2013-11-08	WY
0209709	2005-05-20	NO_LOCATION	WY92WEPR	2011-10-07	WY92WEPR	RSREPAIR	2013-11-08	RSR
0209711	2011-10-07	WY91WEPR	RSREPAIR	2013-11-08	RSREPAIR	WY174VNCR		
0209714	2003-12-15	NO_LOCATION	RSREPAIR					
0209720	2011-10-07	WY95WEPR	RSREPAIR	2013-06-25	RSREPAIR	WY70ASPR		
0209722	2011-10-07	WY106WEPR	RSREPAIR	2013-06-27	RSREPAIR	WY144BSUSR		
0209728	2011-10-07	WY94WEPR	RSREPAIR	2013-11-08	RSREPAIR	WY143NWCPR		
0209729	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN12ASRA	2014-04-04	DN1
0209737	2005-01-11	NO_LOCATION	DN15NWCRB	2006-03-21	DN15NWCRB	RSREPAIR	2006-03-31	RSR
0209739	2011-10-07	WY144WEPR	RSREPAIR	2013-12-09	RSREPAIR	WY178TPR		
0209740	2011-10-07	WY143WEPR	RSREPAIR	2012-09-12	RSREPAIR	DNSPARE	2014-05-30	DN9
0209741	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN10BHR	2014-09-05	DN1

an oil readings dataset (wide version)

```
## # A tibble: 612 × 17
##     Ident                Date WorkingAge   TakenBy    Fe    Al    Cu
##     <chr>               <dttm>      <dbl>     <chr> <dbl> <dbl> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917    13     5    14
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917    18     6    25
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375    26     6    35
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917    15     9    14
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375    14     4    19
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917    18     5    23
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375    24     8    25
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375    27     9    34
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375    16     8    17
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917    20     8    20
## # ... with 602 more rows, and 10 more variables: Cr <dbl>, Si <dbl>,
## #   Pb <dbl>, Ph <dbl>, Ca <dbl>, Zn <dbl>, Mg <dbl>, Mo <dbl>,
## #   Sn <dbl>, Na <dbl>
```

oil readings with Ident and TakenBy properly treated

```
## # A tibble: 612 × 17
##     Ident                Date WorkingAge   TakenBy    Fe    Al    Cu
##    <fctr>              <dttm>      <dbl>    <fctr> <dbl> <dbl> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917    13     5    14
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917    18     6    25
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375    26     6    35
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917    15     9    14
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375    14     4    19
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917    18     5    23
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375    24     8    25
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375    27     9    34
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375    16     8    17
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917    20     8    20
## # ... with 602 more rows, and 10 more variables: Cr <dbl>, Si <dbl>,
## #   Pb <dbl>, Ph <dbl>, Ca <dbl>, Zn <dbl>, Mg <dbl>, Mo <dbl>,
## #   Sn <dbl>, Na <dbl>
```

oil readings dataset (long version)

```
## # A tibble: 7,956 × 6
##     Ident                Date WorkingAge   TakenBy element   ppm
##    <fctr>              <dttm>      <dbl>    <fctr>   <chr> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917      Fe    13
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917      Fe    18
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375      Fe    26
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917      Fe    15
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375      Fe    14
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917      Fe    18
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375      Fe    24
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375      Fe    27
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375      Fe    16
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917      Fe    20
## # ... with 7,946 more rows
```

a (simulated) “gas pipeline” dataset

```
## # A tibble: 1,000 × 4
##      Leak  Size Material Pressure
##    <fctr> <ord>   <fctr>   <fctr>
## 1      No  1.75  Aldyl A     High
## 2      No  1.75  Aldyl A      Med
## 3      No     1  Aldyl A      Low
## 4     Yes   1.5    Steel      Med
## 5      No     1    Steel     High
## 6     Yes     1    Steel     High
## 7     Yes  1.75  Aldyl A      Low
## 8      No  1.75    Steel      Med
## 9      No   1.5  Aldyl A     High
## 10     No  1.75    Steel     High
## # ... with 990 more rows
```

important questions

- where did the data come from?

important questions

- where did the data come from?
  - were the units chosed randomly from a population?

important questions

- where did the data come from?
  - were the units chosed randomly from a population?
  - were the units randomly assigned into groups?

important questions

- where did the data come from?
  - were the units chosed randomly from a population?
  - were the units randomly assigned into groups?
- what are the (joint) *distributions* of the data?

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

Sometimes data come from an *experiment* where units are randomly assigned to different *levels* of one or more *factors*, in which cause cause-and-effect can be inferred using laws of probability.

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

Sometimes data come from an *experiment* where units are randomly assigned to different *levels* of one or more *factors*, in which cause cause-and-effect can be inferred using laws of probability.

Often the data are just some records of what happened. Grander inferences might be made, but only on a subject-matter basis.

distribution (informally)

- A *distribution* is a

distribution (informally)

- A *distribution* is a
  - Complete description of...

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically

distribution (informally)

- A *distribution* is a
  - Complete description of. . .
  - . . . the possible values of one or more variables. . .
  - . . . and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically
  - graphically

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically
  - graphically
- We can also consider probabily models for one or more variables or a relationship among variables. (Focus of this course.)

important concepts from probability

independence

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

$P(A) = 1/2$ and $P(B) = 1/3$, so $P(A)P(B) = 1/6$.

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

$P(A) = 1/2$ and $P(B) = 1/3$, so $P(A)P(B) = 1/6$.

Also, $A \cap B = \{2\}$ so $P(A \cap B) = 1/6 = P(A)P(B)$

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

$P(A) = 1/2$ and $P(B) = 1/3$, so $P(A)P(B) = 1/6$.

Also, $A \cap B = \{2\}$ so $P(A \cap B) = 1/6 = P(A)P(B)$

Conclude: A and B are independent (short form: $A \perp B$.)

independence - definition and example

Two *events* A and B are *independent* if:

$$P(A \cap B) = P(A)P(B),$$

(where $\cap$ means *and*.)

For example, roll a fair die. Let $A = \{1, 2, 3\}$ and $B = \{2, 4\}$.

$P(A) = 1/2$ and $P(B) = 1/3$, so $P(A)P(B) = 1/6$.

Also, $A \cap B = \{2\}$ so $P(A \cap B) = 1/6 = P(A)P(B)$

Conclude: A and B are independent (short form: $A \perp B$.)

Exercise: if $C = \{2, 4, 6\}$ then $B \perp C$ but $A \not\perp C$

independence - comments

Independence is normally something that is *assumed* and not something that is demonstrated.

independence - comments

Independence is normally something that is *assumed* and not something that is demonstrated.

Undisciplined use of language (e.g. "A has nothing to do with B") is the leading cause of error. Use the definition.

independence - comments

Independence is normally something that is *assumed* and not something that is demonstrated.

Undisciplined use of language (e.g. "A has nothing to do with B") is the leading cause of error. Use the definition.

The opposite of independent is "not independent." (Avoid "dependent", which has misleading connotations.)

independence - comments

Independence is normally something that is *assumed* and not something that is demonstrated.

Undisciplined use of language (e.g. "A has nothing to do with B") is the leading cause of error. Use the definition.

The opposite of independent is "not independent." (Avoid "dependent", which has misleading connotations.)

$$A \perp B \iff A \perp B^c \iff A^c \perp B \iff A^c \perp B^c$$

random variables and distributions

concept of random variable

A *random variable* is a rule that assigns a number to any outcome of a random process.

Example: "Roulette". There are 38 slots on a wheel coloured as follows:

Colour	# of slots	Slot labels
Green	2	0, 00
Red	18	1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36
Black	18	2, 4, 6, 8, 10, 11, 13, 15, 17, 20, 22, 24, 26, 28, 29, 31, 33, 35

roulette - II

If bet $100 on "Red", then these are the possibilities:

Result	I receive
Red	200
Not Red	0

Stated another way, here is my net "gain", which I will call X, after the play:

Result	X
Red	100
Not Red	-100

roulette - III

Technically the random variable is this the *rule*:

$$X(1) = X(3) = X(5) = \cdots = X(36) = 100$$
$$X(00) = X(0) = X(2) = \cdots = X(35) = -100$$

roulette - III

Technically the random variable is this the *rule*:

$$X(1) = X(3) = X(5) = \cdots = X(36) = 100$$
$$X(00) = X(0) = X(2) = \cdots = X(35) = -100$$

But this is often a useless technicality. This is all we care about:

x	$P(X = x)$
100	18/38
-100	20/38

This table is the *distribution* of X, i.e. the possible outcomes and their probabilities.

distribution and independence

The distribution of a random variable X is, roughly, all information about the values of X and their probabilities.

distribution and independence

The distribution of a random variable X is, roughly, all information about the values of X and their probabilities.

There's the odd (or maybe not?) fact that when X is *continuously measured* then we have $P(X = x) = 0$ for any particular x. In this case we're concerned with intervals of values and not particular values.

distribution and independence

The distribution of a random variable X is, roughly, all information about the values of X and their probabilities.

There's the odd (or maybe not?) fact that when X is *continuously measured* then we have $P(X = x) = 0$ for any particular x. In this case we're concerned with intervals of values and not particular values.

X and Y can be independent when *knowing the outcome of X does not change the distribution of Y - a very strong statement (usually assumed when appropriate.)

expected value

Random variables can have expected values (averages, means), variances, and standard deviations, that follow these rules:

$$E(aX + b) = aE(X) + b$$

expected value

Random variables can have expected values (averages, means), variances, and standard deviations, that follow these rules:

$$E(aX + b) = aE(X) + b$$
$$E(X + Y) = E(X) + E(Y)$$

expected value

Random variables can have expected values (averages, means), variances, and standard deviations, that follow these rules:

$$
\begin{aligned}
E(aX + b) &= aE(X) + b \\
E(X + Y) &= E(X) + E(Y) \\
\mathrm{Var}(aX + b) &= a^2 \mathrm{Var}(X)
\end{aligned}
$$

expected value

Random variables can have expected values (averages, means), variances, and standard deviations, that follow these rules:

$$
\begin{aligned}
E(aX + b) &= aE(X) + b \\
E(X + Y) &= E(X) + E(Y) \\
\mathrm{Var}(aX + b) &= a^2\mathrm{Var}(X) \\
\mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \text{ when } X \perp Y
\end{aligned}
$$

normal distributions and the central limit theorem

Normal distributions are an important family of symmetric, bell-shaped distributions, parametrized by mean μ and standard deviation σ.

normal distributions and the central limit theorem

Normal distributions are an important family of symmetric, bell-shaped distributions, parametrized by mean μ and standard deviation σ.

They are so widely used *in statistics* because the distribution of a sample average will be approximately normal if the sample size is "large enough".

normal distributions and the central limit theorem

Normal distributions are an important family of symmetric, bell-shaped distributions, parametrized by mean μ and standard deviation σ.

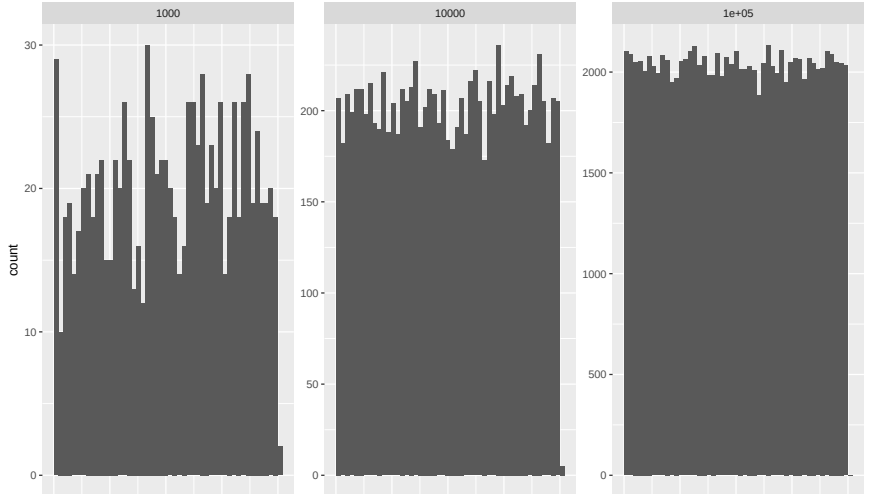
They are so widely used *in statistics* because the distribution of a sample average will be approximately normal if the sample size is "large enough".

"Large enough" is not fixed, but depends on the shape of the undelying population distribution, with more skewness requiring a larger sample size.

normal approximation illustration through simulation - I

I can simulate picking numbers uniformly at random between 0 and 1.

Here are histograms of 1000, 10000, and 100000 picks:



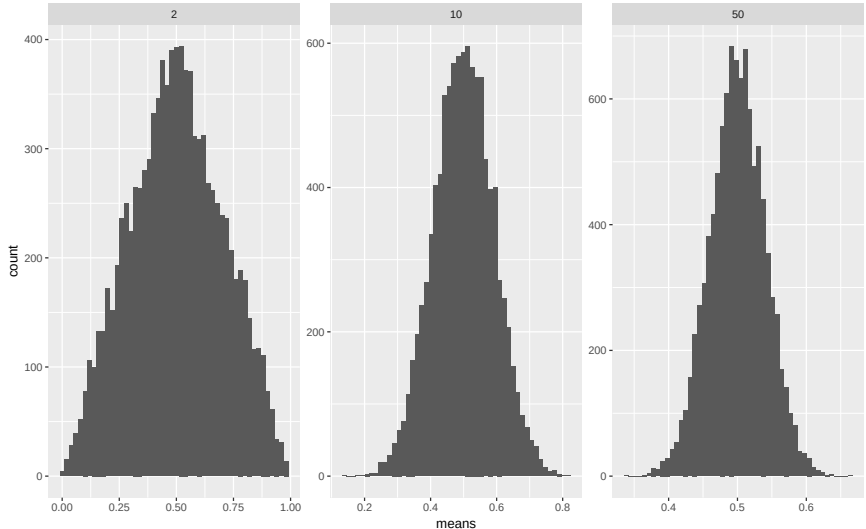
normal approximation illustration through simulation - II

I'll settle on $k = 10000$ "replications" of my simulation.

My simulation will actually consist of: * picking n numbers uniformly at random * calculating the average of those n numbers * doing this k times * making a histogram of the results.

I will choose n to be 2, 10, and 50.

normal approximation illustration through simulation - III



t distributions

If a population is being modeled with a $N(\mu, \sigma)$ probability model and you are going to gather a sample $X_1, X_2, \ldots, X_n$, then the following are true:

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

t distributions

If a population is being modeled with a $N(\mu, \sigma)$ probability model and you are going to gather a sample $X_1, X_2, \ldots, X_n$, then the following are true:

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

We usually don't know σ, but we can estimate it from the data using s, but then:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim t_{n-1}$$